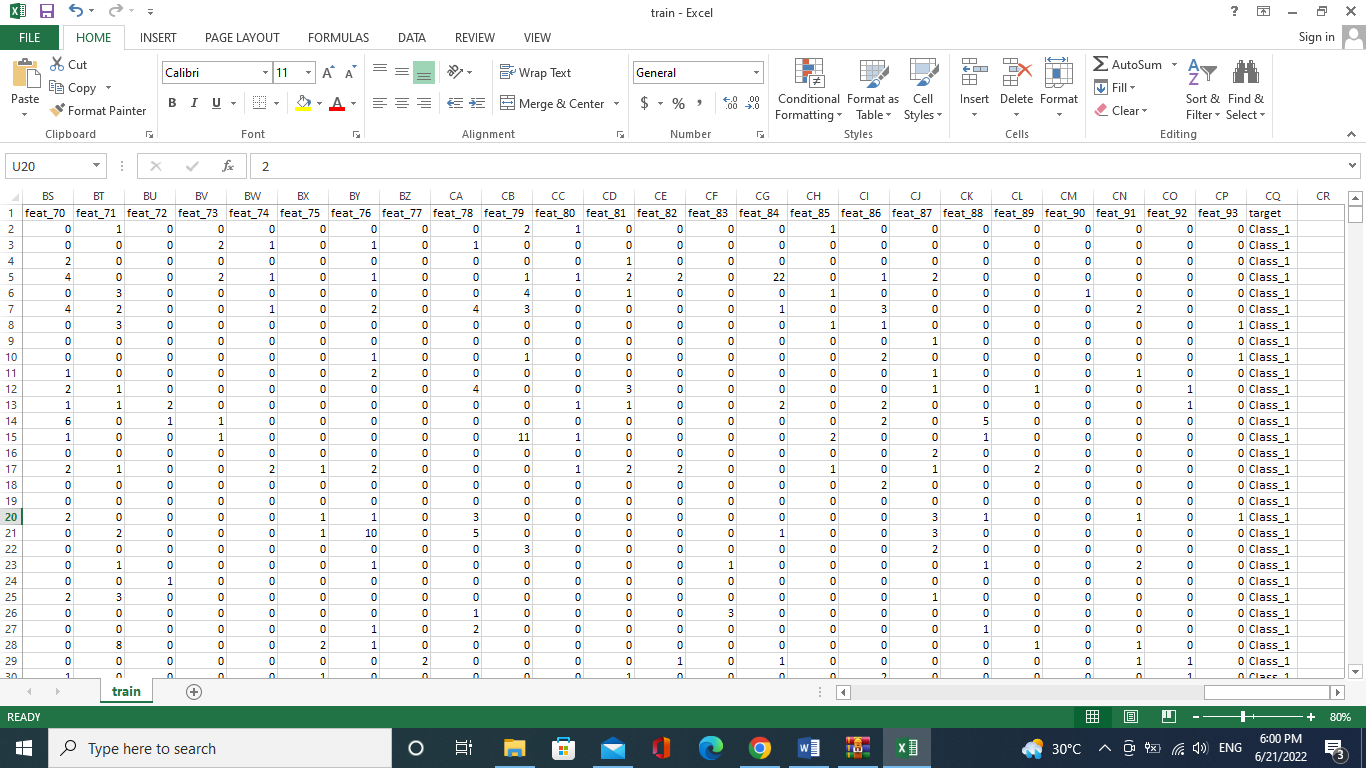
REPORT WEEK 1

**-Ứng dụng mà nhóm chọn tăng tắc** : Training a Machine Learning model (XGBoost)

**-Input**: Đầu vào nhóm sẽ dùng dataset oto trên kaggle <https://www.kaggle.com/c/otto-group-product-classification-challenge/data>

Dataset này bao gồm khoảng 94.000 sản phẩm và 93 input featuresđược chia thành 10 nhóm (ví dụ: thời trang, điện tử, ...). Mục tiêu là xây dựng một model để phân loại một sản phẩm mới vào các nhóm này. Cuộc thi này đã kết thúc vào 05/2015 và người chiến thắng cũng sử dụng XGBoost để tạo model.

-

**-Output**: Model sau khi train của XGBoost.

**-Ý nghĩa của ứng dụng trong thực tế** :

XGBoost là viết tắt của Extreme Gradient Boosting. Đây là thuật toán state-of-the-art nhằm giải quyết bài toán supervised learning cho độ chính xác khá cao bên cạnh mô hình Deep learning.

Vấn đề với các mô hình thường là dữ liệu lớn quá sẽ ảnh hưởng đến thời gian chạy,tốn nhiều chi phí,...

XGboost có tốc độ huấn luyện nhanh, có khả năng scale để tính toán song song trên nhiều server, có thể tăng tốc bằng cách sử dụng GPU, nhờ vậy mà Big Data không phải là vấn đề của mô hình này.

-Ứng dụng này cần tăng tốc.

**-Phát biểu bài toán**:

y là biến ngẫu nhiên “output”.

\mathbf{x} = \{x_1, ..., x_n\} là biến ngẫu nhiên “input” hay “explanatory”.

\{y_i, \mathbf{x}_i\} là mẫu dữ liệu “training”.

F^*(\mathbf{x}) là hàm mục tiêu ánh xạ \mathbf{x} sang y.

L(y, F(\mathbf{x})) là loss function:

* Squared-error: (y - F)^2.
* Absolute error: |y - F|, y \in R^1 (regression).
* Negative binomial log-likelihood: log(1 + e^{-2yF}), y \in \{-1, 1\} (classification).

Mục tiêu của chúng ta tìm được hàm mục tiêu F^* sao cho cực tiểu hoá kỳ vọng của hàm lỗi.

Đặt:

* n: số lượng mẫu huấn luyện.
* m: số lượng features.
* \mathcal{D} = \{(\mathbf{x}_i, y_i)\} là tập dữ liệu với |\mathcal{D}| = n,\mathbf{x}_i \in \mathbb{R}^m, y_i \in \mathbb{R}.
* q: cấu trúc của một cây, ánh xạ mẫu dữ liệu vào nút lá tương ứng.
* T: số lượng nút lá trên cây.
* f_k: cấu trúc các cây k độc lập của mô hình.
* w_i: trọng số của nút lá thứ i.
* \hat{y}_i^{(t)}: giá trị dự đoán của instance thứ i tại vòng lặp thứ t.
* f_t^2(\mathbf{x}_i): đạo hàm bậc 2 của hàm f.
* I_j = \{i|q(\mathbf{x}_i) = j\}: tập các giá trị tại nút lá j
* I_L: tập giá trị nút lá bên trái.
* I_R: tập giá trị nút lá bên phải.
* I = I_L \cup I_R.

**Mô hình học:**

\hat{y}_i = \phi(\mathbf{x}_i) = \Sigma_{k=1}^K f_k(\mathbf{x}_i), f_k \in \mathcal{F}. Trong đó, \mathcal{F} = \{f(\mathbf{x}) = w_{q(\mathbf{x})}\} (q : \mathbb{R}^m) \rightarrow T, w \in \mathbb{R}^T.

Hàm học:

\mathcal{L}(\phi) = \Sigma_i l(\hat{y}_i, y_i) + \Sigma_k \Omega(f_k). Trong đó, \Omega(f) = \gamma T + \frac{1}{2} \lambda ||w||^2

Tiến trình học:

\mathcal{L}^{(t)} = \Sigma_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t)

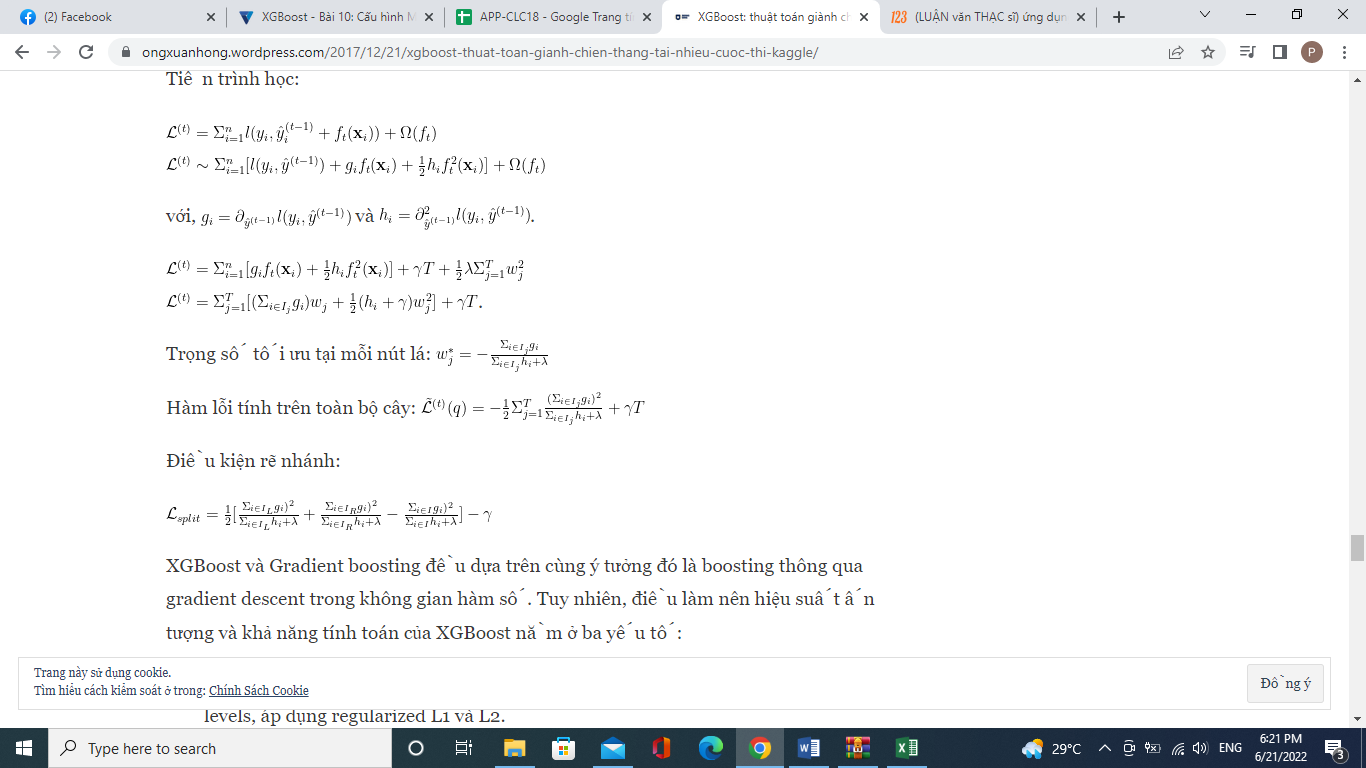
với, g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}) và

h_i = \partial_{\hat{y}^{(t-1)}}^2 l(y_i, \hat{y}^{(t-1)}).

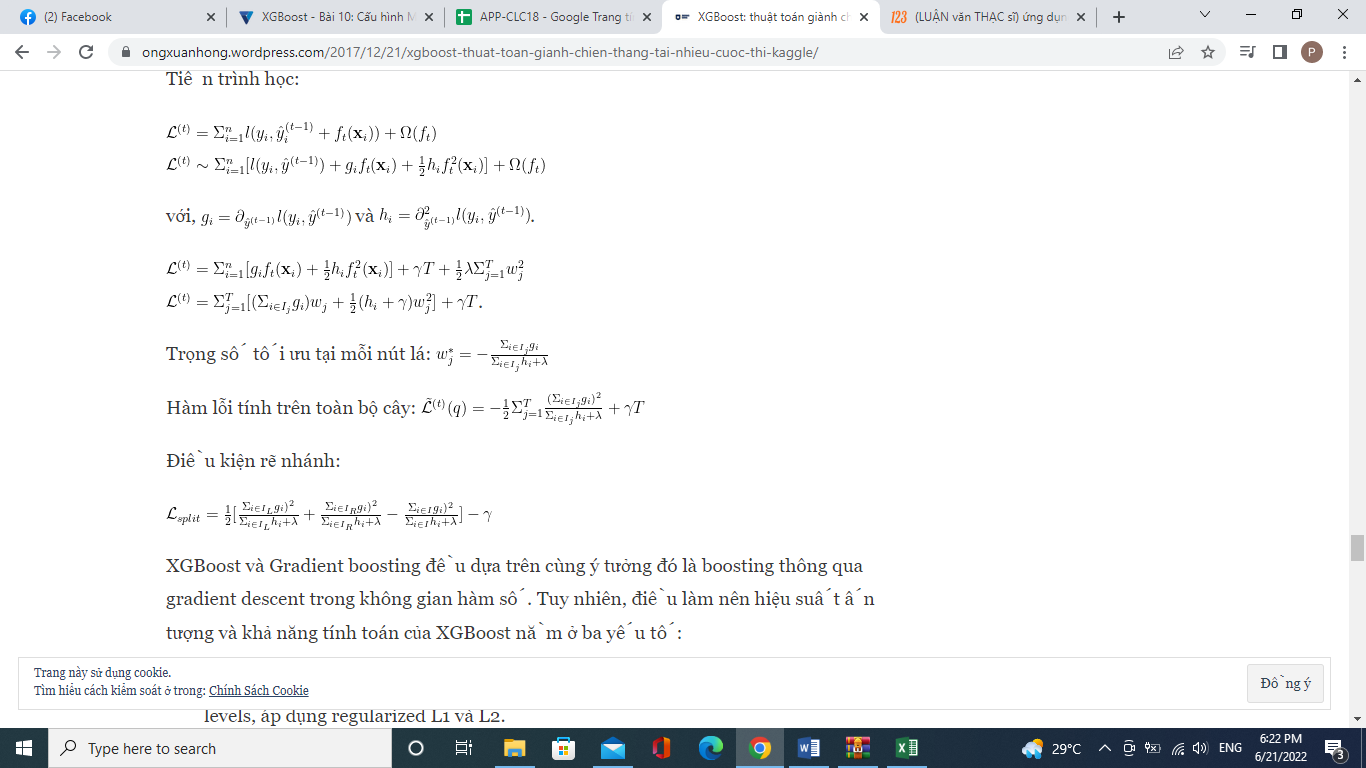
Trọng số tối ưu tại mỗi nút lá:

w_j^* = -\frac{\Sigma_{i\in I_j} g_i}{\Sigma_{i\in I_j} h_i + \lambda}

Hàm lỗi tính trên toàn bộ cây:



Điều kiện rẽ nhánh:



Link colab chạy thử mô hình XGBoost

https://colab.research.google.com/drive/1chhvJBdn\_R1Cw\_W3wu8YKQcWJGeSc\_Qu?hl=vi#scrollTo=RAIwHarqEqg\_&uniqifier=3